IMPROVED WIRELESS DATA TRANSMISSION USING TIME OUT CONTROL

Background of the Invention

1. Field of the Invention

The present invention relates to communications; more specifically, wireless communications.

2. Description of the Related Art

Most wired and wireless data transfers use a common protocol known as TCP/IP (Transmission Control Protocol/Internet Protocol). TCP involves a transmitter sending IP packets of data to a receiver that replies with acknowledgment messages (ACKs) when data is correctly received. Data packets that are lost or corrupted are not acknowledged and are retransmitted until they are received successfully.

In addition to insuring that all transmitted data is correctly received, the ACK scheme also serves to meter the flow of data. When TCP transmissions begin, the initial transmitted data rate is low and ramps up as ACKs are received through a process called "Slow Start". The Slow Start process begins with the transmitter sending one packet of data and waiting for an acknowledge message from the receiver. After receiving the first ACK, the transmitter then sends two packets of data. Each time an ACK is received in response to a transmission, the number of packets transmitted in the next transmission is increased resulting in an exponential increase in data rate illustrated as Ramp 10 in FIG. 1. Eventually a point is reached when , the maximum data rate of the transmission path is attained (point 12 in FIG. 1) and the transmitter sends data at a rate equal to the rate the receiver acknowledges it.

Transmission at the maximum data rate will be persist until an error occurs indicated by failure of the transmitter to receive an acknowledgement of data sent to the receiver. Should this occur the transmitter assumes the data is lost due to network congestion and invokes the Slow Start procedure. This is illustrated in portion 14. It should be noted that the rate at which the data transmissions are ramped up to the maximum transmission rate, is dependent on how quickly acknowledgments arrive at the transmitter and hence on the the round trip delay in the communication channel. For example, each step up in the ramp involves transmitting one or more packets of data and waiting for one or more acknowledgments from a receiver. In a system such as a wired communication system, the round trip delays may be as little as 1 or 2 milliseconds and as a result, the ramp up period is relatively quick. Unfortunately, in wireless communication systems the round trip delay may be on the order of 100 or 200 milliseconds. As a result, the ramp up period in wireless communication systems is relatively long as illustrated by dashed ramp up curve 16. As a result, in wireless communication systems, the slower ramp up time results in a waste of channel capacity illustrated by area 18.

T1 T2 problem

Normally, errors are indicated by repeated acknowledgements with the same sequence number, indicating that the receiver is receiving data but a packet has been lost. To handle cases where many packets are lost, and no acknowledgments are made, TCP also includes a time out period. If an acknowledgment is not received within the time out period the transmitter will assume all unacknowledged packets are lost and begin retransmission and Slow Start. Such timeouts rarely occur in a wired system because the transmission path is reliable and multiple packet losses are very rare. In a wireless system, varying conditions of noise, fading, and channel allocation among multiple users can cause delays in data transmission sufficient to cause TCP timeouts even when no packets are lost. Therefore wireless data are far more vulnerable to time outs than wired systems and their associated slower ramp up of data transmission rates results in a greater reduction of channel capacity per time out.

T3

Summary of the Invention

The present invention decreases the probability of a time out and the resulting waste of channel capacity in a communication system by operating in a manner that elevates the length of the time out period so as to minimize the number of time outs that occur. In the current version of TCP (dubbed RENO and used by almost all computers on the Internet) the length of the time out is based on the sum of the average channel round trip delay and four times the deviation from the average in the channel round trip delay. A delay is introduced into the communication channel so as to increase the deviation from the average in the channel delay. This results in an increase in the length of time required for a time out. As a result, the number of time outs is drastically decreased which in turn decreases the number of wasteful ramp up times that results in a more efficient use of channel capacity.

T4

Brief Description of the Drawings

FIG. 1 illustrates data transfer rate vs. time for TCP data transmission;

FIG. 2 illustrates an example of a wireless channel's bandwidth as a function of time;

FIG. 3 illustrates a functional block diagram of a wireless communication channel;

FIG. 4 illustrates variations in channel delays as a function of time; and

FIG. 5 illustrates a bimodal distribution of channel delay as a function of time.

Detailed Description of the Invention

FIG. 2 illustrates the bandwidth of a wireless transmission channel over time. It can be seen that dropouts 30 occur from time to time. These dropouts may be a result of fading, noise or sharing the wireless communication channel among several users. For example, each dropout period 30 may be the time during which another user is granted access to the wireless communication channel. In order to minimize the number of bandwidth wasteful ramp up periods, it is desirable to insert sufficient delays into the wireless communication channel so that that the TCP time out period is greater than most of the dropout periods 30.

Sub A1

~~FIG. 3 illustrates a functional block diagram of a wireless communication channel. Base station 40 receives data from a data source 50 hosting an application using TCP. Base station 40 communicates the data over an air interface to mobile station 60 which passes the received data to a data receiver 70 hosting an application using TCP. The delay can be inserted into the communication channel at either base station 40 or mobile station 60. In base station 40, data is transmitted and received via RF section 80. Channel delay can be inserted into data being transmitted by base station 40 or into acknowledges received by base station 40. Delays may be inserted data transmitted by base station 40 using buffer 82. Buffer 82 may be a shift register or cyclically addressed memory. Processor 84 controls the delay by controlling the number of stages the data must pass through when passing through buffer 84. Processor 84 monitors the channel delay by monitoring acknowledge messages received from RF section 80. As a result, processor 84 can modify the depth or amount of delay added by buffer 82 until the desirable delay is measured as seen by the delay in acknowledges received in response to data transmissions.~~

It is also possible to add delay by using buffer 86 to delay the acknowledge messages passed from base station 40 to application 50. Once again, buffer 86 may be a variable length or stage buffer such as a shift register or cyclically addressed memory. By adjusting the delay provided by buffer 86, processor 84 increases the channel delay as seen by application 50.

Whether the delay is provided using buffer 82 or buffer 86, application 50 is made to see longer delays which result in longer TCP time out periods.

In a similar fashion, mobile station 60 may also add delay to the transmission channel. Data is transmitted from and received through RF section 90 of mobile station 60. Processor 92 may control channel delay by controlling the depth of outgoing data buffer 94 or it may control the depth of acknowledge buffer 96. In either case, application 70 sees a greater channel delay which results in a longer TCP time out period. It should be noted that periods of other communication protocols may be controlled in a similar manner, and the approach outlined applies regardless of whether 50 transmits and 70 receives or 70 transmits and 50 receives.

FIG. 4 illustrates channel delay as a function of time. It can be seen that the channel delay is clustered around an average delay. Additionally, it should be noted that the deviation or the absolute value of the averaged distance between the different data transmission rates and the average transmission rates is rather small. Equation 1 illustrates that the time out (T_0) as defined by TCP is the sum of the average channel delay plus four times the deviation of the channel delays.

$$t_{ave} + 4\ t_{dev} = T_0 \qquad \text{EQ. 1}$$

FIG. 5 illustrates channel delays as a function of time where delay is added to the channel in order to control and thereby increase the length of the TCP time out. Assuming the same channel as the channel of FIG. 4, approximately 50 percent of the channel transmissions are not delayed which results in a cluster of channel delays along line 100, which is the average channel delay when no additional delays are added to the channel. Approximately, 50 percent of the transmissions are delayed to produce a second cluster of channel delays along line 102. This produces a larger deviation and results in an average channel delay illustrated by line 104. It should be noted that by delaying approximately 50 percent of the transmissions, a bimodal or substantially bimodal delay distribution is achieved. A bimodal distribution is desirable because is maximizes the deviation for a given increase in average delay time. Using Equation 1 it can be seen that the new time out associated with the delay pattern of FIG. 5, results in a time out (T_0^2) that is many times larger than the prior time out of T_0^1. As a result, by providing a relatively small increase in the average channel delay, the time out used by TCP is increased dramatically. It is desirable to add delays so that a time out length (T_0) is created that is larger than the typical dropout in the wireless communication channel bandwidth.

It should be noted that delay may be added at either the base station or mobile station and it may be added using one or both of the data transmission paths or the acknowledge receive path.

It is also possible to control delay by scheduling multiple users to use the same communication channel. This can be accomplished by initially providing each user with a short amount of time using the communication channel. The initial short amount of time should be short enough that a time out does not occur while at the same time increasing the delay in the communication channel per user so that the time out used by the TCP protocol is increased.